REMARKS

This Amendment is filed in response to the Office Action dated December 22, 2004. All objections and rejections are respectfully traversed.

Claims 1-23 are in the case.

Claims 19-23 have been added to better claim the invention.

Claim 14 has been amended to better claim the invention.

The Title of the invention has been amended to correct a typographical error. No new matter has been entered, and the Title is believed to be in allowable condition.

Rejections under 35 U.S.C. §112, Second Paragraph

At page 2 of the Office Action, claims 1, 2, 7, 8, 13, and 14 were rejected under 35 U.S.C. §112, second paragraph, as being indefinite for failing to particularly point out and distinctly claim the subject matter which applicant regards as the invention. Particularly, claims 1, 7, and 13 recite "the server's *capacity* to handle connections," and claims 2, 8, and 14 recite the term "*quotient/remainder*."

Applicant directs the Examiner to the "Background of the Invention" of the present Application, where Applicant states:

> For instance, the method sends a new connection to a server which has the lowest metric, wherein the metric is defined as the number of connections on the server divided by the weight (*or capacity*) of the server. This metric is kept as a *quotient/remainder* pair. To keep track of the metric and

> the remainder, integer division is typically performed on all servers every time a connection is added or removed. (Application page 2, lines 13-17, *emphasis added*)

Applicant respectfully urges that "the server's *capacity* to handle connections" is generally known to those skilled in the art, particularly with respect to load balancing algorithms.

Applicant directs the Examiner to the enclosed documentation as representatives of terms used by those skilled in the art. For instance, the enclosed Glossary from the "Catalyst 4840G Software Feature and Configuration Guide" (available from www.cisco.com) defines the following:

> **Weighted Least Connection:**
> Load-balancing algorithm in which the next real server chosen for a new connection to the virtual server is the server with the fewest active connections. Each real server is assigned a weight, n, that represents *its capacity to handle connections*, as compared to the other real servers associated with the virtual server. The server with the fewest connections is based on the number of active connections on each server, and on the relative capacity of each server. *The capacity of a given real server* is calculated as the assigned weight of that server divided by the sum of the assigned weights of all of the real servers associated with that virtual server, or n1/(n1+n2+n3...). (Glossary page 3)

The use of assigned weights and server *capacity* is further described in the enclosed document, "Cisco – Understanding CSM Load Balancing Algorithms" at printout page 3, under "Weighted Round Robin and Weighted Least Connections." Illustratively, this document states: "Issue the weight command … to configure *the capacity of the real servers* in relation to the other real servers in the server farm."

Applicant also respectfully urges that the term "*quotient/remainder*" is generally known to those skilled in the art, particularly with respect to load balancing algorithms.

Applicant directs the Examiner to another enclosed document, "Job Scheduling Algorithms in Linux Virtual Server" that demonstrates the division necessary to result in the "*quotient/remainder*" pair in question. Specifically, Applicant points out the section labeled "Weighted Least-Connection Scheduling" that spans printout pages 3-4, that discusses in detail how "[t]he Virtual Server Administrator can assign a weight to each real server, … in which the percentage of the current number of live connections for each server is a ratio to its weight." More particularly, this document explains the ratio as the number of alive connections for a server divided by the server's weight (C/W). Notably, as mentioned above, Applicant defines a metric as the "number of connections on the server divided by the weight (*or capacity*) of the server." This metric is thus stored as "a *quotient/remainder* pair," in that it is the result of an integer division, with a quotient and remainder (as their accepted meanings).

Applicant respectfully urges, therefore, that "the server's *capacity* to handle connections," and the term "*quotient/remainder*" are both generally known to those skilled in the art, and that claims 1, 2, 7, 8, 13, and 14, accordingly, are believed to be in condition for allowance.

Rejections under 35 U.S.C. §102(a)

At page 4 of the Office Action, claims 1, 7, and 13 were rejected under 35 U.S.C. §102(a) as being anticipated by He et al., U.S. Patent No. 6,671,259, issued on December 30, 2003, hereinafter He.

The present invention, as set forth in representative claim 1, comprises in part:

> 1. A method for load balancing a plurality of servers, the method comprising:
>
> providing a plurality of control blocks, each control block associated with a number of active connections a server is connected with, the control block configured to control at least one server with the associated number of connections in a server list;
>
> causing each control block to point to a server with a least value ascertained by determining the number of connections on the server relative to the server's capacity to handle connections;
>
> ***selecting the control block associated with the least number of connections; and***
>
> ***selecting the server pointed to by the control block.***

He discloses a method and system for load balancing a network having a plurality of client systems and servers, where a load balancing server receives client requests and selects an appropriate server to balance tasks among the plurality of servers. Namely, the selection in He is based on network measurement (e.g., the total amount of current client requests, or load) performed on each of the servers or gathered from each server using network measurement devices and techniques known to those skilled in the art (Column 4, lines 5-17). He characterizes (or categorizes) servers based on those network measurements, such as into high load or low load servers, and selects the optimal server for an incoming request.

Applicant respectfully urges that He does not show Applicant's claimed novel "***selecting the control block associated with the least number of connections; and selecting the server pointed to by the control block.***"

Applicant claims a method and system for load balancing a plurality of servers by using a plurality of novel control blocks. These control blocks each represent a number of active connections (e.g., 0, 1, 2, 3, etc.), and each of the plurality of servers is associated with an appropriate control block. For instance, a server with one connection is associated with the control block representing one connection, a server with two connections is associated with the control block representing two connections, etc. Each server associated with a control block is then ranked according to a value related to the number of connections relative to the server's capacity to handle connections, where the control block points to the server with the least value. For example, where first and second servers have the same number of connections, and the first server has a capacity that is higher than that of the second server, the first server will have a lower value, and will be pointed to by the control block. From this arrangement, Applicant load balances the servers by simply selecting the control block with the least number of connections, and selecting the server that is pointed to by the control block (i.e., with the lowest value).

He does not address the use of control blocks. He obtains network measurements such as the number of client requests or load directly from the servers, and bases its calculations for the optimal server on these obtained measurements. Each time the system in He receives a client request, a load balancing server must examine the network load measurements obtained from the servers to select the most optimal server (Column 4, lines 22-24). In contrast, Applicant's claimed control blocks obviate the need to examine network load measurements from the servers each time a client request is received. Instead, Applicant simply selects the most optimal server to service the request based on the arrangement of the servers within the control blocks.

Applicant respectfully urges that the He patent is legally precluded from anticipating the claimed invention under 35 U.S.C. §102 because of the absence from the He patent of Applicant's claimed "***selecting the control block associated with the least number of connections; and selecting the server pointed to by the control block***."

Rejections under 35 U.S.C. §103(a)

At page 5 of the Office Action, claims 2, 3, 5, 8, 9, 11, 14, 15, and 17 were rejected under 35 U.S.C. §103(a) as being unpatentable over He in view of Yu et al., U.S. Patent No. 6,078,943, hereinafter Yu. Also, at page 7 of the Office Action, claims 4, 6, 10, 12, 16, and 18 were rejected under 35 U.S.C. §103(a) as being unpatentable over He in view of Yu, in view of Ernst, U.S. Patent No. 6,298,371. Claims 2-6, 8-12, and 14-18 are dependent claims that are believed to be dependent from allowable independent claims, and therefore in condition for allowance.

All independent claims are believed to be in condition for allowance.

All dependent claims are believed to be dependent from allowable independent claims, and therefore in condition for allowance.

Favorable action is respectfully solicited.

Please charge any additional fee occasioned by this paper to our Deposit Account No. 03-1237.

Respectfully submitted,

James M. Behmke

James M. Behmke
Reg. No. 51,448
CESARI AND MCKENNA, LLP
88 Black Falcon Avenue
Boston, MA 02210-2414
(617) 951-2500